Predicting isoform transcripts: What does the comparison of known transcripts in human, mouse and dog tell us?

Nicolas Guillaudeux, Catherine Belleannée, Samuel Blanquart, Jean-Stéphane Varré

Predicting isoform transcripts: What does the comparison of known transcripts in human, mouse and dog tell us?

Nicolas Guillaudeux[1], Catherine Belleannée[1], Samuel Blanquart[1], Jean-Stéphane Varré[2]

[1] Dyliss team, Univ Rennes, Inria, CNRS, IRISA, Rennes F-35000, France.
[2] Univ. Lille, CNRS, Centrale Lille, UMR 9189 – CRIStAL, F-59000 Lille, France.

nicolas.guillaudeux@inria.fr

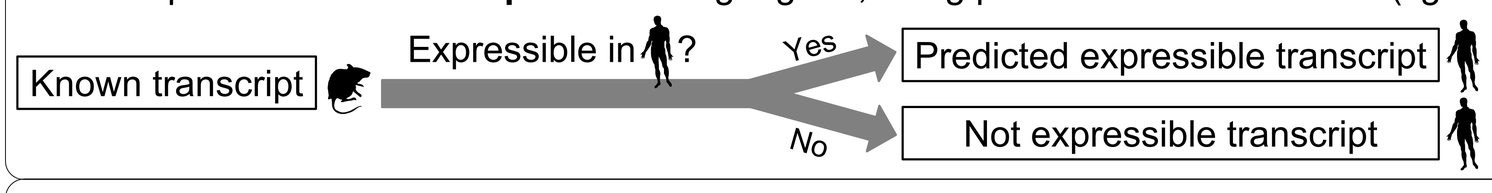
To fill in the knowledge about **transcript isoforms** expressed from a gene, we have proposed a **comparative genomics** method allowing to identify orthologous exons shared by a pair of genes [1]. We predict transcript isoforms in human, mouse and in a non-model organism, dog, and we identify **135 conserved genes** having common gene structures and common potential transcriptomes.

Modelling gene structures using comparative genomics to predict isoforms [1]

Our **structure of a** target **gene** (fig.1a)

- based on functional sites of known transcripts:
  [: start codon **]** : stop codon **< >** : splice sites **a letter**: (part of) a coding exon

- using functional sites of known transcripts in an orthologous source gene (fig.1b):
  - ➢ to reveal **new functional sites** or **coding exons** on the target gene (fig.1c)
  - ➢ to predict **new transcripts** of the target gene, using predicted functional sites (fig.1d)



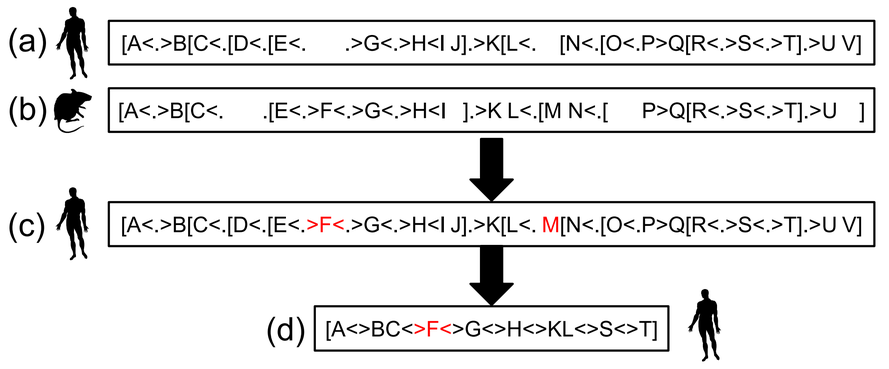


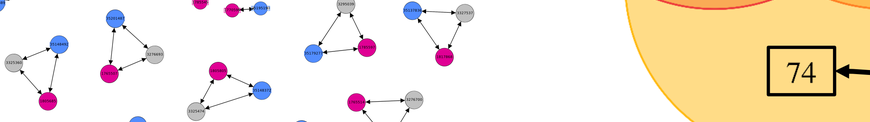
Figure 1. CREM gene: pairwise comparison between human (a) and mouse (b) reveals new functional sites or exons (red) on human gene (c). This lead to predict new expressible transcripts (d) in human.

Application: Predicting transcripts in human, mouse and dog

Data used:
2,167 orthologous genes & 18,109 known transcripts:
- human and mouse*: CCDS*
- dog*: ENSEMBL*

Pairwise comparisons [1] and merging results

6,861 new predicted transcripts (fig. 2):



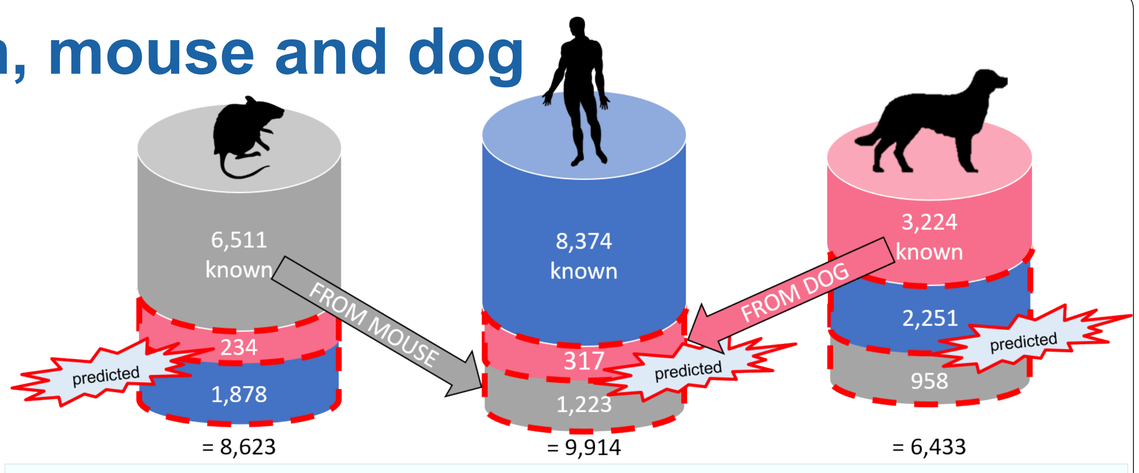


Figure 2. Estimated transcriptomes: known and predicted transcripts obtained using pairwise comparisons.

Analyzing conservation over functional sites and transcripts

We build two conservation graphs for each gene:

- conserved **functional site** graph (fig.4) & conserved **transcript** graph (fig.5)
- a graph component shows orthology relationships between species
- classification of graph components (3-species case):
  - shared in the 3 species (fig.3a), in 2 species (fig.3b), specific to a species (fig.3c)
- only graphs without “ambiguous” components (fig.3d) are **considered for analysis**

Graph components:

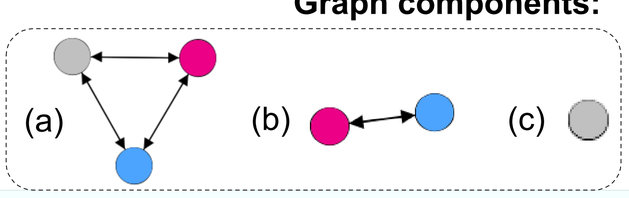


Figure 3. Considered gene components: (a) triplet, (b) duplicate, (c) singleton. “Ambiguous” graph components (d) are not considered.

Conservation of functional sites:
each node is a functional site of a gene
each edge represents an orthology relationship between two functional sites

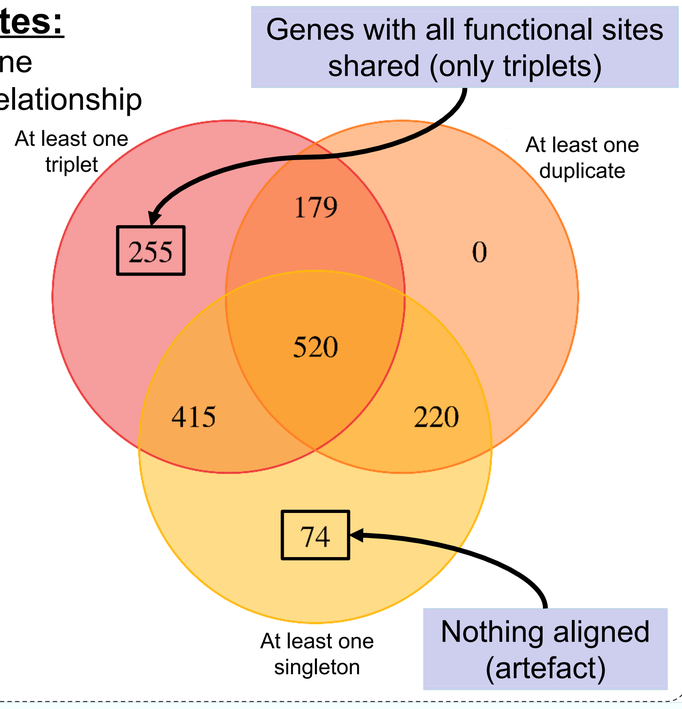
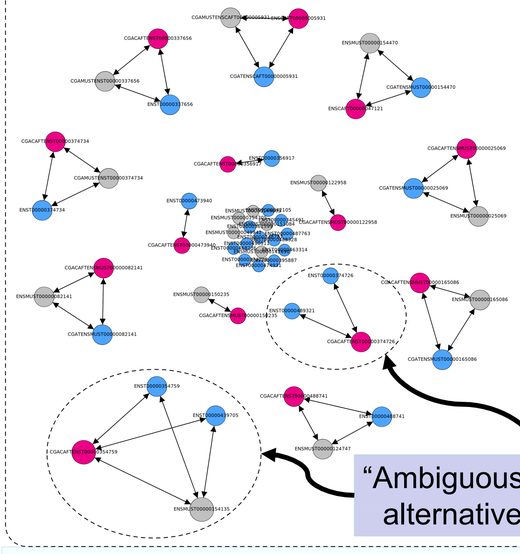


Figure 4. Functional site graph of CREM gene (left). Analysis of **1,663 considered** functional site graphs. 255 graphs reveal genes having common gene structures (right).

Conservation of transcripts:
each node is a transcript of a gene
each edge represents an orthology relationship between two transcripts



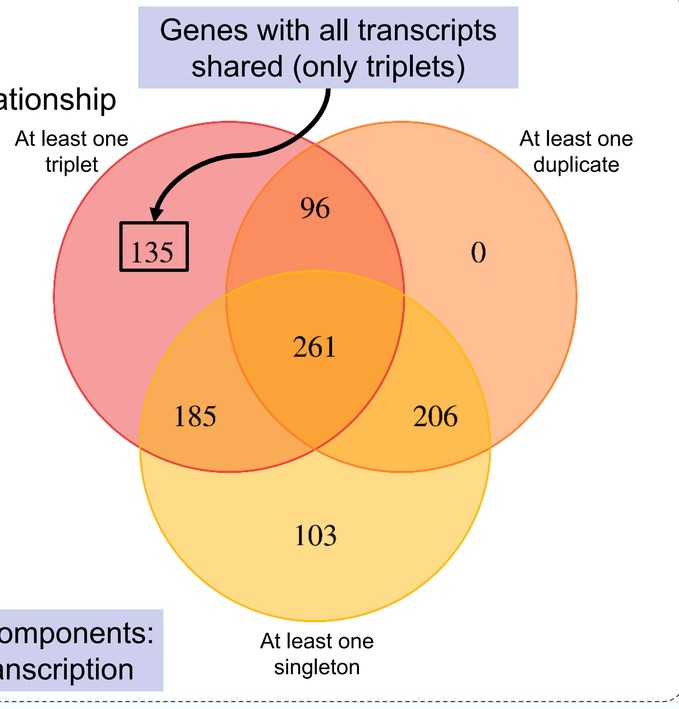
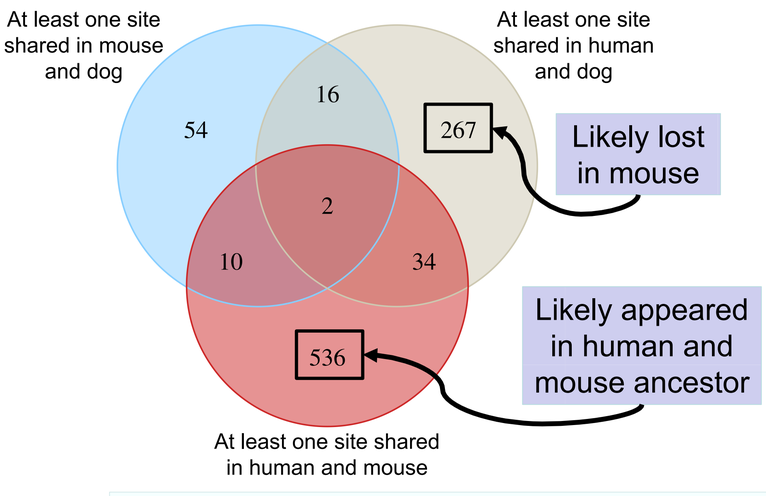


Figure 5. Transcript graph of CREM gene (ambiguous) (left). Analysis of **986 considered** transcript graphs. 135 graphs reveal genes having common potential transcriptomes (right).

Phylogenetic interpretation of gene structures

Most analyzed genes encountered structure divergence



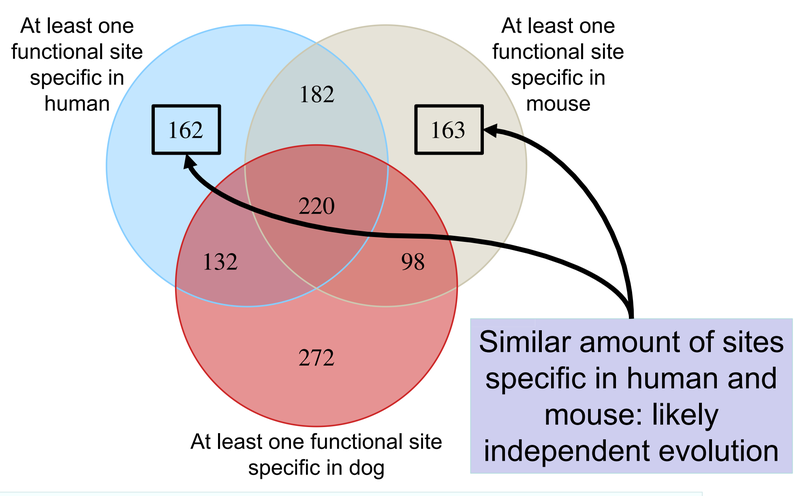
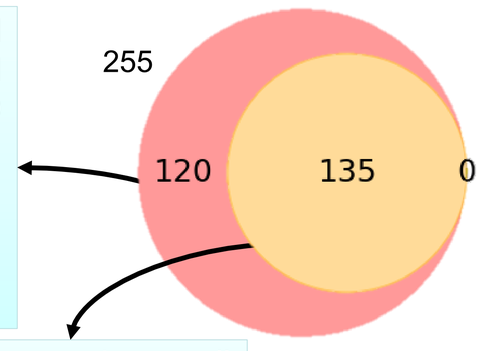


Figure 6. Distribution of duplicates (left) and singletons (right) components in functional site graphs.

Some genes have the same gene structure but different transcriptomes

255 genes have all functional sites conserved over the 3 species but 120 express different transcripts. Example:
- Frameshifts
- Alternative transcription
- False negative



135 genes have all functional sites and transcripts conserved over the 3 species

[1] S. Blanquart et al., “Assisted transcriptome reconstruction and splicing orthology.”, BMC Genomics 17 (2016).